SUPPLEMENTARY METHODS

Leif Väremo, Jens Nielsen and Intawat Nookaew
December 10, 2012

Simulated p-value distributions

The distribution of p-values can be modeled by a mixture of the beta distribution and the uniform distribution (1). We simulated five p-value distributions, with increasing skewedness towards low p-values, by generating 6000 random numbers using the uniform distribution and a varying amount (100, 500, 1000, 1500 and 2500) of random numbers using the beta distribution (with α=0.02 and β=1). For each of the five simulated distributions we randomly generated 100 gene sets for each size between 1 and 1000 genes, in total yielding 100,000 gene sets. For these gene sets the gene set p-values were calculated using the standard Fisher's and Stouffer's methods to show their dependence on size.

Details on the gene set statistics methods

This section will describe, for each gene set statistic, what gene-level statistics they take and how they handle them, as well as which significance assessment approaches are possible to run and which directionality classes that will be calculated. But first, a general note is that the sample permutation approach cannot be used together with subsets of the gene-level statistics. This is due to the fact that, for a given gene set, the number of up-regulated and down-regulated genes, i.e. the sizes of the subsets, will be varying between permutations. Thus, the background distribution of gene set statistics would be based on varying sizes, which would highly impact the results.

Fisher's combined probability test

Given a gene set i, containing n genes with p-values $p_1, p_2, \dots, p_n$, the gene set statistic, S_i, is calculated according to (2):

$$S_i = 2 \sum_{j=1}^{n} -\ln p_j \ .$$

For a gene set with n genes, the significance can be calculated by:

$$p_{S_i} = 1 - \chi^2(S_i; 2n)$$

where $\chi^2(x; df)$ is the (non-central) chi-squared distribution function with df degrees of freedom and the non-centrality parameter set to zero. The significance can also be estimated by gene sampling or sample permutation. Fisher's method can result in the mixed-directional and non-directional classes. Using all unmodified p-values will result in the non-directional class, and gene sets with a low p_{S_i} are interpreted as significantly changed gene sets (there can be a mix of up- and down-regulated genes, as long as they as a group are significantly changed). If subsets of up- and down-regulated genes are handled separately, two p_{S_i} (of the mixed-directional class) will be generated, which we can call $p_{S_i,mix,up}$ and $p_{S_i,mix,dn}$. Gene sets with a

low $p_{S_i,mix,up}$ will be significantly affected by up-regulated genes, disregardful of the extent of down-regulated genes, and the reverse for $p_{S_i,mix,dn}$.

Stouffer's method

Given a gene set i, containing n genes with p-values $p_1, p_2, \ldots, p_n$, the gene set statistic, S_i, is calculated according to (3):

$$S_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \theta^{-1}(1 - p_j),$$

where θ^{-1} is the inverse normal cumulative distribution. The gene set p-value can then be calculated by using the normal cumulative distribution function:

$$p_{S_i} = 1 - \theta(S_i).$$

The significance can also be estimated by gene sampling or sample permutation. Stouffer's method can result in all directionality classes. The mixed-directional and non-directional classes are calculated in the same way as described for the Fisher's method. For the distinct-directional class the original gene p-values are transformed in order to contain information about the direction of change. This will result in the two one-sided gene p-values described below:

$$p_{j,up} = \begin{cases} p_j/2 & \text{if gene } j \text{ is up-regulated} \\ 1 - p_j/2 & \text{if gene } j \text{ is dn-regulated} \end{cases}$$

$$p_{j,dn} = 1 - p_{j,up}$$

A gene set will in this case contain p-values for all its genes, independent of direction, but significant genes in one direction, e.g. up-regulated, will have low p-values, whereas significant genes in the other direction, e.g. down-regulated, will have p-values close to one. By using these two gene-level p-value types with the Stouffer's method, two gene set p_{S_i} will be returned, which we call $p_{S_i,distinct,up}$ and $p_{S_i,distinct,dn}$. Now, a gene set with a low $p_{S_i,distinct,up}$ can be interpreted as being significantly affected by up-regulation, but not a mix of up- and down-regulation. The opposite is true for $p_{S_i,distinct,dn}$. If using gene-level p-values, for the distinct-directional class, the most negative effects on the gene set p-value will come from significant genes of the opposite direction, whereas the most negative effect for the non-directional class will come from genes that are non-significant. This means that not necessarily all distinct-directional significant gene sets will be non-directional significant.

Reporter features

Given a gene set i, containing n genes with p-values $p_1, p_2, \ldots, p_n$, the gene set statistic, S_i, is calculated according to (4):

$$S_i = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \theta^{-1}(1 - p_j),$$

where θ^{-1} is the inverse normal cumulative distribution. This is identical to Stouffer's method. For significance estimation, the gene set statistic is normalized with the mean and standard

deviation of the background gene set statistics for that particular gene set. The background gene set statistics are calculated by randomizing the genes labels and recalculating the gene set statistic many (e.g. 10 000) times. The gene set p-value can then be calculated back from the normalized Z-scores, by using the normal cumulative distribution function,

$$p_{S_i} = 1 - \theta\left(\frac{S_i - \mu_n}{\sigma_n}\right).$$

The reporter algorithm can also be used with gene sampling or sample permutation for calculating gene set p-values. In this case it will be identical to running Stouffer's method in the same manner. The reporter features algorithm can result in all directionality classes, see Stouffer's method for details.

Parametric Analysis of Gene Set Enrichment

The PAGE gene set statistic (5) is calculated according to the following equation:

$$S_i = \frac{M - \mu}{\sigma}\sqrt{n},$$

where μ and σ are the mean and standard deviation of all the gene fold changes (or similar gene statistics, we use the gene t-values), and M is the mean of the fold change values of the n genes in the gene set. PAGE only results in the distinct-directional class. The cumulative normal distribution can be used to calculate distinct-directional p-values from the gene set statistics:

$$p_{\text{distinct,up}} = 1 - \theta(S_i)$$

$$p_{\text{distinct,dn}} = \theta(S_i).$$

Gene sampling or sample permutation can also be used to calculate the gene set p-values. In this case the background fraction that is smaller than the gene set statistic is used for $p_{\text{distinct,dn}}$ and the opposite for $p_{\text{distinct,up}}$.

Tail strength

First the p-values of the genes belonging to gene set i are ordered so that $p_1 \leq p_2 \leq \cdots \leq p_n$. The gene set statistic is calculated according to (6):

$$S_i = \frac{1}{n}\sum_{j=1}^{n}\left(1 - p_j\frac{n+1}{j}\right)$$

The gene set p-values are calculated using either gene sampling or sample permutation. The tail strength method can result in all directionality classes, see Stouffer's method for details.

Wilcoxon rank-sum test

The Wilcoxon rank sum test is equivalent to a Mann-Whitney test. The gene set statistic is calculated according to the implementation of the Wilcoxon rank sum test in R using the function wilcox.test. If r_j is the rank (where rank 1 is given to the smallest number) of gene statistic j among all genes and gene set i contains n genes, the gene set statistic can be calculated from the sum of the ranks of the genes in the set according to:

$$S_i = \sum_{j=1}^{n} r_j - \frac{n(n+1)}{2}.$$

For large sample sizes (>50), the distribution of a rank sum, W, can be assumed to follow a normal distribution with mean μ and standard deviation σ:

$$\mu = \frac{n(N-n)}{2},$$

$$\sigma = \sqrt{\frac{n(N-n)(N+1)}{12}},$$

where n is the number of genes in the gene set and N is the total number of genes. It follows that

$$\text{pr}(W \geq S_i) \approx \text{pr}\left(Z \geq \frac{S_i - \mu}{\sigma}\right),$$

where $Z \sim \text{Normal}(0,1)$. The Wilcoxon rank-sum test can result in gene set p-values of all classes. In practice, the arguments alternative="less" or alternative="greater" are used with the wilcox.test function, testing if the ranks of the genes in the gene set are smaller than the ranks of all the other genes, or the opposite, respectively. Alternatively, gene sampling or sample permutation can be used to estimate gene set significance. If gene-level p-values are used, the handling is the same as for e.g. Stouffer's method to produce the different p-value classes. For t-values, the absolute values are used for the non-directional class. For the mixed-directional class, the up- and down-subsets are handled separately and as absolute values. For the distinct-directional class, the t-values are kept as they are.

Gene set enrichment analysis

The GSEA method (7,8) takes t-like statistics, s_j, as input and ranks them so that $s_1 \leq s_2 \leq \cdots \leq s_N$. Next, a running sum is computed, starting with the first statistic and moving to the last. For each statistic, if it belongs to the gene set (hit), the running sum is increased, and if the statistic does not belong to the gene set (miss), the running sum is decreased. The enrichment score, the gene set statistic, is the maximum deviation from zero of the running sum. Formally:

$$P_{hit}(S,i) = \sum_{\substack{s_j \in S \\ j \leq i}} \frac{|s_j|^p}{N_S}, \quad \text{where} \quad N_S = \sum_{s_j \in S} |s_j|^p,$$

$$P_{miss}(S,i) = \sum_{\substack{s_j \notin S \\ j \leq i}} \frac{1}{N-n},$$

where N is the total number of genes, n is the number of genes in the gene set, and p is an exponent parameter (defaulting to 1). The enrichment score statistic is the maximum deviation of $P_{hit}(S,i) - P_{miss}(S,i)$ from zero, for $l = 1, \ldots, N$. If a gene set contains genes in the extreme ends of the list, e.g. with either low or high ranks, corresponding to high positive or negative gene statistics, the gene set will get a high enrichment score. The GSEA method only results in the distinct-directional class. The GSEA method calculates the gene set p-values using either gene sampling or sample permutation but uses the positive and negative background

portion separately for significance calculation. If a gene set statistic is negative, the gene se p-value is the fraction of the negative background portion that is smaller than the gene set statistic. If the statistic is positive, the p-value is the fraction of the positive background portion that is larger than the statistic. As a consequence of this, a given gene set can only have one of $p_{distinct,up}$ and $p_{distinct,dn}$, the other is set to NA. These p-values are the same as GSEA is referring to as nominal p-values.

The GSEA method also contains a normalized enrichment score (NES) and false discovery rate (FDR). Although these steps are not addressed in this paper, they are implemented in the Piano package. A normalized enrichment score, NES, is calculated by dividing each positive enrichment score with the average of its positive background permutation enrichment scores, i.e. the positive part of the background scores for that gene set. Similarly, a negative enrichment score is divided by the average negative background permutation enrichment score. If there is no positive (or negative) background, the enrichment score is divided by zero. Next, all the background enrichment scores are also normalized in the same manner, i.e. gene set by gene set. An FDR is calculated for each gene set with a positive NES according to:

$$\mathrm{FDR}_{\mathrm{up}} = \frac{(\#\ \text{background NES} \geq \mathrm{NES}^{*})/\ (\#\ \text{background NES} \geq 0)}{(\#\ \mathrm{NES} \geq \mathrm{NES}^{*})/(\#\ \mathrm{NES} \geq 0)}$$

where NES^{*} is the NES for the selected gene set, the background is in this case the background for all gene sets together, and NES are for all gene sets. Similarly, an FDR is calculated for each gene set with a negative NES.

Mean, median and sum

The gene set statistics are simply calculated as the mean, median or sum of the gene-level statistics. All directionality classes can be calculated. For the non-directional class, if t-values are used, the absolute values of the t-values are used in order to find gene sets that are significantly changed, disregardful of direction. F-values and p-values are kept as they are. For the mixed-directional class, subsets of the gene-level statistics are used correlating to up- and down-regulated genes. If t-values are used, the absolute values are used, so that only positive scores are achieved for both the up and down gene set p-values. For the distinct-directional class the gene t-values are kept as they are. A gene set, containing genes that are equally up- and down regulated, will have a gene set statistic close to zero, but if the genes are mostly regulated in one direction, the gene set statistic will be either large positive or negative. If gene-level p-values are used, they are handled in the same way as described for Stouffer's method in order to return the distinct-directional class. Gene sampling or sample permutation is used for estimating gene set significance.

Maxmean statistic

For gene set i the gene set statistic is the maximum of the absolute averages of the negative and positive parts of the gene statistics belonging to the set (9):

$$S_i = \max\left(\bar{s}_i^{(+)}, -\bar{s}_i^{(-)}\right),$$

$$\bar{s}_i^{(+)} = \frac{1}{n}\sum_{j \in i} s_j^{(+)},$$

$$\bar{s}_i^{(-)} = \frac{1}{n}\sum_{j \in i} s_j^{(-)},$$

where $s_j^{(+)}$ is the j:th positive gene-level statistic in the gene set, $s_j^{(-)}$ is the j:th negative gene-level statistic, and n is the total number of genes in the set. Since the maxmean method is defined to detect gene sets that are strongly regulated in either or both directions, only the non-directional p-values can be calculated. The most negative effect on the gene set p-values, is from non-significant genes regulated in both directions. However, if a gene set is significantly up-regulated, the gene set statistic will be unaffected by the magnitude of non-significant down-regulated genes. Gene sets that are significant contain genes that are either significantly up-regulated, significantly down-regulated, or both. This differs slightly from the above methods that also return non-directional p-values, since they find gene sets that are significantly regulated, but will not find gene sets in which a small portion of genes are significantly up but a large portion of genes are down but non-significant. Such a gene set will be found by the maxmean method.

Robustness analysis of the consensus scoring

For the robustness analysis, the GSA workflow was rerun for the human diabetes data with the 149 Mootha gene sets (7) using a selection of methods (Fisher's method, Stouffer's method, Reporter features, PAGE, Tail strength, mean, median, sum and Maxmean) using different numbers of gene permutations (500, 750, 1000, 1500, 2000, 5000, 10000). The results from the runs in each of the seven permutation groups were aggregated so that each group resulted in a consensus score vector for each directionality class for the gene sets. These consensus score vectors were compared by calculating the Spearman correlation between all possible pairs of the vectors. To reflect the worst case, the minimum of these pairwise correlations was recorded. In the end, we thus have a minimum correlation for each directionality class in each of the four rank aggregation methods (mean, median, Copeland and Borda). As a summary the minimum correlation over all directionality classes is reported for each of the four aggregation methods.

Next, to test the consistency of the result when mixing GSA runs with different numbers of permutations, each GSA run was randomly chosen from one of the seven permutation groups, resulting in a set of runs with mixed permutation numbers. These runs were aggregated and consensus scores were calculated by each of the four rank aggregation methods and this procedure was repeated 1000 times. Similarly as above, the minimum Spearman correlation between the 1000 consensus score vectors and over all the directionality classes is reported for each rank aggregation method.

Secondly, the robustness with regard to which GSA runs are used as input to the consensus scoring algorithm was investigated. This was done using the same GSA results

that are discussed in the main paper, i.e. those presented in Figure 2, 3 and 5, using all combinations of settings for the *Saccharomyces cerevisiae* data (with GO terms) and the human diabetes data (with GO terms and the 149 Mootha gene sets). The analysis was performed by randomly selecting 95% of the available GSA runs, for a given dataset, as input (in principle 95% of the gene set p-value vectors for a given directionality class) and calculating the consensus scores with each of the four rank aggregation methods. This was repeated 1000 times, each time randomly selecting 95% of the runs. In a similar fashion as above, the minimum pairwise Spearman correlation of the 1000 consensus score vectors and for all directionality classes was calculated (in total 5000 runs of the consensus scoring algorithm with varying input), representing the worst case for each rank aggregation method. This approach was also repeated while using 85% and 70% of the GSA runs.

The above robustness analyses are on a global scale, i.e. the correlation is based on all gene sets. In order to also focus on the top-ranked gene sets an additional approach was performed. Here, we collected the gene sets that were among the top 10% in at least one of the repeated runs of the consensus scoring algorithm using the median rank as consensus score (e.g. for the 1000 runs calculating the consensus scores for the non-directional class when randomly using 95% of the GSA results as input, all gene sets that were ranked among the top 10%, in at least one of the 1000 runs, were collected in a list). Next, we report the fraction of these gene sets that are among the top 10% and among the top 20% in all consensus scoring runs. For a satisfactory result, these fractions, in particular the latter, should be high.

1. Allison, D.B., Gadbury, G.L., Heo, M., Fernández, J.R., Lee, C.K., Prolla, T.A. and Weindruch, R. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.*, **39**, 1-20.
2. Fisher, R.A. (1932) Statistical methods for research workers. Oliver and Boyd, Edinburgh.
3. Stouffer, S.A., Suchman, E.A., Devinney, L.C., Star, S.A. and Williams Jr, R.M. (1949) The American soldier: adjustment during army life. Princeton University Press, Oxford, England.
4. Patil, K.R. and Nielsen, J. (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 2685-2689.
5. Kim, S.Y. and Volsky, D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinf.*, **6**, 144.
6. Taylor, J. and Tibshirani, R. (2006) A tail strength measure for assessing the overall univariate significance in a dataset. *Biostatistics*, **7**, 167-181.
7. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M. and Laurila, E. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267-273.
8. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545-15550.
9. Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107-129.